TESTING GROUPED DATA FOR EXPONENTIALITY
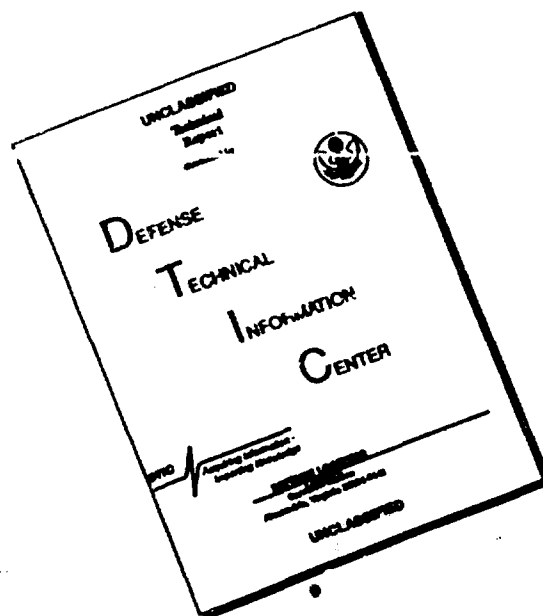
Ernest M. Scheuer

SEP 9 1968

The RAND *Corporation*

SANTA MONICA · CALIFORNIA

CLEARINGHOUSE

39

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

MEMORANDUM
RM-5692-PR
AUGUST 1968

TESTING GROUPED DATA FOR EXPONENTIALITY

Ernest M. Scheuer

The RAND *Corporation*

1700 MAIN ST · SANTA MONICA · CALIFORNIA · 90406

PREFACE

This Memorandum derives from RAND's continuing interest in
statistics and data analysis. It will be useful to computer-model
builders and others who, on the basis of grouped data, wish to test
the hypothesis that some event time is described by an exponential
distribution.

The study was suggested by a RAND colleague who wanted to
determine whether a given empirical distribution intended for use
in the SAMSOM computer model could be more simply described by an
exponential distribution.

SUMMARY

This expository Memorandum treats the problem of testing grouped data for fit to an exponential distribution, provides a JOSS[†] computer program to implement an appropriate test, and gives examples showing how to use this program. The Memorandum also contains a brief discussion of testing nongrouped data for exponentiality and of testing grouped data for fit to certain nonexponential distributions.

This is not a mathematical treatise. No proofs are given. Even so, if a reader finds a particular discussion too mathematical for his tastes, he can skip over it and not miss the essential parts of this Memorandum.

[†]JOSS is the trademark and service mark of The RAND Corporation for its computer program and services using that program.

ACKNOWLEDGMENTS

CONTENTS

BLANK PAGE

1. INTRODUCTION

Certain event times (such as the time-to-failure of some types of equipment, the length of telephone calls, the time between the emission of α-particles from a radioactive source, etc.) are thought to be described by an exponential distribution. This Memorandum discusses the "chi-square minimum" procedure for testing the hypothesis of exponentiality when the data are event frequencies in intervals of the time axis. That is, the individual event times are not available, only the number of events observed in each of a set of mutually exclusive intervals--where these intervals taken together cover the entire nonnegative time axis. Data arise in this form, for example, when, rather than recording individual occurrence times, only the number of occurrences in certain time periods (days, hours, minutes, etc.) are noted. There is, however, no requirement that the various periods be of equal length.

To be explicit about the problem, let k-1 nonnegative numbers, $T_1 < \ldots < T_{k-1}$, be given. These numbers divide the nonnegative time axis into k mutually exclusive and exhaustive intervals:[†]

$$0 \leq t < T_1, \; T_1 \leq t < T_2, \; \ldots, \; T_{k-2} \leq t < T_{k-1}, \; t \geq T_{k-1} \;.$$

The data are the number of events observed in each interval: f_1 in $0 \leq t < T_1$, f_2 in $T_1 \leq t < T_2$, $\ldots$, f_k in $t \geq T_k$. This Memorandum discusses the chi-square minimum test of the hypothesis that these data are

---

[†]The intervals need not be closed on the left and open on the right as given here. There must, however, be no overlap among the intervals, and every point of the nonnegative time axis must be in some interval.

consistent with an exponential distribution and provides a JOSS computer program to implement this test. Worked examples are given. In addition, there is a brief discussion of the modified chi-square minimum test of exponentiality for grouped data, of tests of exponentiality for nongrouped data, and of tests of fit of grouped data to other distributions than the exponential.

2. THE CHI-SQUARE AND CHI-SQUARE MINIMUM TESTS[+]

For convenience in the subsequent discussion, denote the interval $0 \leq t < T_1$ by I_1, ..., the interval $T_{j-1} \leq t < T_j$ by I_j, ..., and the interval $t \geq T_{k-1}$ by I_k. Recall that the data available are the number of observations f_j in interval I_j, $j=1$, .., k. Denote by n the total number of observations, $\sum_1^k f_i$.

2.1 THE CHI-SQUARE TEST OF A COMPLETELY SPECIFIED HYPOTHESIS

Consider first the situation in which one wishes to test the hypothesis H that the data come from a completely specified probability distribution function. That is, the probability content p_j of the interval I_j is calculable under the hypothesized distribution. A measure of the discrepancy between the sample and the hypothesized distribution is the chi-square expression

$$\chi^2 = \sum_{j=1}^{k} [(f_j - n_i {}_j)^2 / (np_j)] \tag{2.1}$$

The asymptotic distribution of χ^2 is the chi-square distribution with k-1 degrees of freedom. Thus one rejects the hypothesis H if χ^2 is larger than the critical value of chi-square for k-1 degrees of freedom at the significance level being used. This assumes that n is sufficiently large so that the distribution of χ^2 is essentially at

---

[+]A general reference for this section is Cramér (1951), Chap. 30.

its asymptotic distribution. Some authors suggest that the expected

frequencies, np_j, be 10 or more; others say at least 5--combining

intervals, if necessary, to achieve this condition. Cochran (1954)

suggests that these rules are far too conservative and that expecta-

tions as low as one may be used in the tails of a unimodal distribution.

Suppose one wants to test the hypothesis that the data--intervals

I_j containing f_j observations, $j=1, \ldots, k$--are consistent with the

exponential distribution with occurrence rate ρ. This distribution

has density function

$$f(t) = \begin{cases} \rho \exp(-\rho t) & , \quad t \geq 0 \\ 0 & , \quad t < 0 \end{cases} \qquad (2.2)$$

so that, writing $T_o = 0$ and $T_k = + \infty$, one has

$$p_j = \int_{T_{j-1}}^{T_j} \rho \exp(-\rho t) \, dt$$

$$= \exp(-\rho T_{j-1}) - \exp(-\rho T_j), \text{ for } j=1, \ldots, k . \qquad (2.3)$$

These probabilities p_j must be substituted in Eq. (2.1) to calculate
x^2.

The situation discussed in the preceding paragraph is *not* the

one of primary interest in this study. The concern here is with

testing the hypothesis that the data come from *some* exponential dis-

tribution, and not that they come from a *given* exponential distribu-

tion. Section 2.2 discusses the latter situation, first in general

terms, and then specializes to the case of the exponential distribu-

tion.

2.2 THE CHI-SQUARE MINIMUM METHOD

As before, K intervals, I_1, ..., I_k, are given that are mutually exclusive and exhaustive for the random variable under study. Data consist of the number of observations f_1, ..., f_k, respectively, in these intervals. The hypothesis to be tested is that the data come from a probability distribution that assigns probability content $p_j(\alpha_1, ..., \alpha_s)$ to the interval I_j. The function $p_j(\alpha_1, ..., \alpha_s)$ is of known form, but involves s($<$ k-1) unknown parameters, α_1, ..., α_s. These parameters are associated with the hypothesized distribution function. If α_1, ..., α_s were known, then it would only be necessary to calculate the probabilities $p_j(\alpha_1, ..., \alpha_s)$, substitute them for p_j in Eq. (2.1), and follow the procedure of the preceding subsection. When α_1, ..., α_s are unknown, the following procedure--called the chi-square minimum method--can be used. Choose the values α_1, ..., α_s that minimize the expression

$$\chi^2(\alpha_1, ..., \alpha_s) = \sum_{j=1}^{k} [(f_j - np_j(\alpha_1, ..., \alpha_s))^2/(np_j(\alpha_1, ..., \alpha_s))] . \quad (2.4)$$

The minimizing values, α_1', ..., α_s', say, are called the chi-square minimum estimates of α_1, ..., α_s. The asymptotic distribution of $\chi^2(\alpha_1', ..., \alpha_s')$, the minimum value of $\chi^2(\alpha_1, ..., \alpha_s)$, is the chi-square distribution with k-s-1 degrees of freedom. The estimates α_1', ..., α_s' and $\chi^2(\alpha_1', ..., \alpha_s')$ can be found by a numerical minimizing technique or by solving for α_1, ..., α_s the system of equations that result from equating to zero the partial derivative of Eq. (2.4) with respect to α_i, i=1, ..., s. This system of equations is

$$\sum_{j=1}^{k} \left[\frac{f_j - np_j(\alpha_1, \ldots, \alpha_s)}{p_j(\alpha_1, \ldots, \alpha_s)} + \frac{(f_j - np_j(\alpha_1, \ldots, \alpha_s))^2}{2np_j^2(\alpha_1, \ldots, \alpha_s)} \right] \frac{\partial p_j(\alpha_1, \ldots, \alpha_s)}{\partial \alpha_i} = 0,$$

$$i = 1, \ldots, s . \qquad (2.5)$$

For the exponential distribution there is only one parameter (i.e., $s = 1$) and it has previously been denoted ρ. The probability content of the interval I_j, $p_j(\rho)$, previously given in Eq. (2.3), is

$$p_j(\rho) = \exp(-\rho T_{j-1}) - \exp(-\rho T_j) , \text{ for } j=1, \ldots, k .$$

The derivative of $p_j(\rho)$ with respect to ρ is given by

$$\frac{dp_j(\rho)}{d\rho} = T_j \exp(-\rho T_j) - T_{j-1} \exp(-\rho T_{j-1}) , \text{ for } j=1, \ldots, k . \qquad (2.6)$$

[Since $T_k = +\infty$, the terms $\exp(-\rho T_k)$ and $T_k \exp(-\rho T_k)$ must be taken to be zero.]

In the exponential case, the equations shown in (2.5) reduce to a single equation. A JOSS program has been written to solve this equation for its root ρ' and to calculate $\chi^2(\rho')$. The program is listed in the Appendix, and details of the program are discussed in Sec. 4. Input requirements are given, and sample problems are worked in Sec. 5.

2.3 SOME SHORTCOMINGS OF CHI-SQUARE PROCEDURES

Cochran (1954) points out that chi-square tests are not directed against any specific alternative to the null hypothesis, that is, to detect any particular pattern of deviations $(f_i - np_i)$ that may hold if

the null hypothesis is false. A consequence is that chi-square tests
are often insensitive and do not indicate significant results when
the null hypothesis is actually false.

Berkson (1966) makes a suggestion along the lines of overcoming
the insensitivity to the pattern of deviations. He recommends that one
compute many tests with different interval widths. The context of
Berkson's remark is one in which the statistician has the interval width
at his disposal. In the situation considered here the intervals are
given, so that one only has the option of combining adjacent intervals to
make wider (but fewer) intervals--not of considering intervals shorter
than those given. Nonetheless, if a particular chi-square test fails
to reject the null hypothesis, combining the data into different classes
and making further chi-square tests may indicate whether the failure
is due to a particular pattern of discrepancy between observation and
hypothesis to which the chi-square test is insensitive, or whether the
observations really conform to the hypothesis.

Epstein (1960) remarks that "the chi-square goodness of fit test
has several drawbacks. Among them are its large sample character and
dependence upon the choice of the number and position of the intervals
into which the time axis is divided." An illustration of this latter
point is given here in Sec. 5, Example 2.

In short, chi-square tests tend to have low power, i.e., relatively
low probability of rejecting a null hypothesis when it is false, and
are sensitive to the arrangement of the data into groups.

Section 7 discusses tests of exponentiality against specified
alternatives when the individual observations--not merely grouped
data--are available.

3. THE MODIFIED CHI-SQUARE MINIMUM METHOD

There is another general approach to testing goodness of fit, a method closely related to the chi-square minimum method, which deserves mention here. This is the so-called "modified chi-square minimum" method (Cramér, 1951, Chap. 30). Cramér states that the system of equations (2.5) is often difficult to solve, and that it can be shown that the influence of the second term in the square brackets in Eq. (2.5) becomes negligible as n becomes large. Neglecting this second term and invoking the condition $\sum_{j=1}^{k} p_j(\alpha_1, \ldots, \alpha_s) = 1$ for each possible s-tuple $(\alpha_1, \ldots, \alpha_s)$ yields the system of equations

$$\sum_{j=1}^{k} \frac{f_j}{p_j(\alpha_1, \ldots, \alpha_s)} \cdot \frac{\partial p_j(\alpha_1, \ldots, \alpha_s)}{\partial \alpha_i} = 0 \ , \ i=1, \ldots, s \ . \ (3.1)$$

It can be shown that the solution of Eqs. (3.1), say, $\alpha_1^*, \ldots, \alpha_s^*$, exists and is unique and that the asymptotic distribution of $\chi^2(\alpha_1^*, \ldots, \alpha_s^*)$ is the same as that of $\chi^2(\alpha_1', \ldots, \alpha_s')$ discussed in Sec. 2, namely, the chi-square distribution with k-s-1 degrees of freedom.

Cramér does not remark on the relative speeds of convergence of the modified chi-square minimum and the chi-square minimum to the limiting distribution (nor am I aware of any other investigation of this question), so there is no reason, on this basis, to prefer one method to the other. If, as Cramér suggests, Eqs. (3.1) are easier to solve than Eqs. (2.5), then this might make the modified chi-square minimum method preferable. It must be remembered, however, that the first printing in America of Cramér's book appeared in 1946 and there

was a printing in Sweden in 1945. *This was well before the computer
era.* It is difficult to see where, with the aid of a computer, Eqs.
(3.1) would be substantially easier to solve than Eqs. (2.5). Both
systems require the probability content of the intervals and the par-
tial derivatives of these probabilities with respect to the parameters.
Eqs. (2.5) do involve a few more arithmetic operations, but not
dramatically more.[†]

3.1 APPLICATION TO TESTING FOR EXPONENTIALITY

The following is an adaptation to the exponential distribution
of a discussion by Cramér (1951, pp. 437-438) for the normal distribution.

As before, the data are frequencies f_i in the interval $[T_{i-1}, T_i)$,
i=1, ..., k. If the hypothesis of exponentiality is true, the prob-
ability p_i corresponding to the i-th class is

$$p_i = \int_{T_{i-1}}^{T_i} \rho \exp(-\rho x) dx \quad . \tag{3.2}$$

[It turns out to be convenient to ignore the fact that this integral
can be expressed in a closed form. Cf. Eq. (2.3).] The derivative
of p_i with respect to ρ , needed in Eq. (3.1), is

$$\frac{dp_i}{d\rho} = \int_{T_{i-1}}^{T_i} (1-\rho x) \exp(-\rho x) dx \quad . \tag{3.3}$$

An approximate solution to Eq. (3.1) can be given (1) if it is possible

---

[†]It is hoped that no reader will consider the foregoing discussion
as a criticism of Cramér's monumental work. We could hardly have
expected Cramér, in the early forties, to have anticipated the power
of modern computer technology.

to arrange the grouping so that there are no observations in the last interval (i.e., $f_k = 0$), and (2) if each interval but the last is sufficiently short so that each integral in Eqs. (3.2) and (3.3) can be approximated by the value of the integrand at the midpoint, ξ_j, of the corresponding interval times the length of that interval. One has

$$\sum \frac{f_j}{P_j} \cdot \frac{dp_j}{d\rho} = \sum \frac{f_j}{(T_j - T_{j-1}) \exp{(-\rho \xi_j)}} \cdot (T_j - T_{j-1})(1 - \rho \xi_j) \exp{(-\rho \xi_j)} = 0 \quad . \quad (3.4)$$

Simple algebra yields as an approximate solution to Eq. (3.4)

$$\hat{\rho} = n/\sum f_j \xi_j \quad . \tag{3.5}$$

This is the estimate for the rate parameter of an exponential distribution calculated from a grouped sample according to the usual rule that all sample values in a certain class are treated as though they were at the midpoint of the class interval. Cf. the discussion accompanying Eq. (4.3).

If the conditions in the sentence following Eq. (3.3) are satisfied, then substituting $\hat{\rho}$ from Eq. (3.5) into Eq. (2.1) will yield a fairly reasonable approximation to the value of χ^2 that would be obtained by either the chi-square minimum or the modified chi-square minimum method.

An illustration of the above approximation is given in Example 4 of Sec. 5.

4. JOSS NUMERICAL PROCEDURES FOR THE CHI-SQUARE MINIMUM TEST OF EXPONENTIALITY

The chi-square minimum procedure requires finding the root, say ρ', of the equation

$$\sum_{j=1}^{k} \left[\frac{f_j - np_j(\rho)}{p_j(\rho)} - \frac{(f_j - np_j(\rho))^2}{2np_j^2(\rho)} \right] \frac{dp_j(\rho)}{d\rho} = 0 . \qquad (4.1)$$

The expressions $p_j(\rho)$ and $dp_j(\rho)/d\rho$ are given by Eqs. (2.3) and (2.6), respectively.

The root ρ' of Eq. (4.1) is obtained in our JOSS program by Newton's method. Although Newton's method (also called the Newton-Raphson method) is discussed in almost every numerical analysis text,[†] a brief description of it follows. To solve the equation $\phi(x) = 0$ by Newton's method, select a starting value x_0 and calculate a sequence of values $\{x_i\}$ by the relation

$$x_{i+1} = x_i - \phi(x_i)/\phi'(x_i) , \ i = 0, 1, 2, \ldots . \qquad (4.2)$$

A criterion that can be used to stop the iteration is to terminate at that i* which yields $|\phi(x_{i*})| < \epsilon$ for a suitably chosen (small) value of ϵ.

This stopping rule, with $\epsilon = 10^{-6}$, is used in our solution of Eq. (4.2). A slight variant of Eq. (4.2) is used in our calculations. Instead of the derivative $\phi'(x)$, the approximating difference quotient $[\phi(x+d) - \phi(x)]/d$, with $d = 10^{-4}$, is used. This is the version of

---

[†]See, for example, Ralston (1965), Sec. 8.4, especially p. 332.

Newton's method given by Bryan and Paxson (1967, p. 5.19).

The right-hand side of Eq. (4.1), with (2.3) and (2.6) plugged in, is denoted M(r) in the JOSS program (r for ρ). Newton's method is applied to M(r) to yield the solution--which is denoted R in the JOSS program. The starting value is taken to be[†]

$$n \left[\sum_{i=1}^{k-1} f_i (T_{i-1} + T_i)/2 + f_k T_{k-1} \right]^{-1} , \qquad (4.3)$$

and this value is chosen for the following reason. If all the individual occurrence times t_1, ..., t_n were available, one could calculate the maximum likelihood estimate of ρ : $n/\sum_1^n t_i$. In the absence of these data, all observations in an interval are treated as though they were at the mid-point of the interval--save for the last interval in which all observations are treated as if they were at the lower limit of that interval. (The last interval is treated differently because its upper limit is infinite.)

_____

[†]Actually the program works with $X_i = T_i/T_{k-1}$ instead of T_i. This is to avoid the possibility of dealing with very large negative powers of e, i.e., very small numbers, in calculating the interval probabilities (2.3). Once the minimizing rate parameter has been calculated with the X's, it is only necessary to divide it by T_{k-1} to get the desired value--the minimizing rate parameter associated with the T's. Of course, this is all internal to the program, and the casual user need not even be aware of this detail.

5. DESCRIPTION OF THE JOSS PROGRAM
AND WORKED EXAMPLES

5.1 PROGRAM DESCRIPTION

To use the program, enter it into JOSS and command JOSS to "Do part 1."[+] JOSS will request k, the number of intervals; $T(1)$, ..., $T(k-1)$, the boundaries between the intervals; and $f(1)$, ..., $f(k)$, the observed frequencies in the intervals. After a pause, during which JOSS solves Eq. (4.1), the output appears. JOSS prints out the interval numbers, the lower and upper boundaries of the intervals (the upper boundary of the final interval, infinity, is denoted inf.), the number of observations in each interval, and the "expected number" of observations in each interval. These latter quantities are equal to $np_j(\rho')$, where $p_j(\rho)$ is given by Eq. (2.3), and ρ' is the solution to Eq. (4.1). These expected frequencies are rounded to two places in the output, but not at all in the calculations. The total, n, of observations is shown, as is the sum of the expected frequencies—this solely as a check. The calculated value of chi-square (the value of Eq. (2.1) with $p_j = p_j(\rho')$) is shown, together with its degrees of freedom.[++]

[+]In reading the following description of JOSS input requirements and output format, the reader may want to look at samples of each in Example 1 on p. 14.

[++]The significance of this value for the indicated number of degrees of freedom must be determined from a table of percentage points of the chi-square distribution. Our JOSS program does not have these percentage points built into it. This should cause no hardship, however, as chi-square tables are widely available.

The chi-square minimum estimate of the rate parameter ρ and of its reciprocal, the mean, are also printed out. The latter values correspond to the rate and mean of the exponential distribution of "best" fit to the data--best in the sense of minimizing chi-square. They are of dubious significance when the hypothesis of exponentiality is rejected, for if a distribution of occurrence times is nonexponential, then the occurrence rate is not constant, but varies with time.[†]

5.2 WORKED EXAMPLES

Example 1: The data in Table 1 constitute unscheduled maintenance net aircraft turnaround times for 237 sorties. After noting that

Table 1

NET TURNAROUND TIMES

Interval (Hours)	Frequency
0-1	9
1-2	46
2-3	48
3-4	23
4-5	15
5-6	24
6-7	8
7-8	15
8-9	7
9-10	4
10-11	5
11-12	4
more than 12	29
Total	237

[†]If $f(t)$ is the density function, and $F(t)$ is the cumulative distribution function, then $r(t)$, the *occurrence rate function*, is defined by $r(t) = f(t)/[1-F(t)]$.

there are 13 intervals, these data are entered into JOSS. Figure 1 is a reproduction of the JOSS interrogation (to the left of the equal sign) and the user response (to the right of the equal sign). The JOSS output is given in Fig. 2. A computed chi-square value of 55.53 for 11 degrees of freedom is significant beyond the 0.1 percent level, so the hypothesis of exponentiality for t.e turnaround time data of Table 1 is decisively rejected. Since the hypothesis of exponentiality is not supported, the estimates of the occurrence rate and the mean should be ignored. (See the last remark in Sec. 5.1.)

$$k = 13$$
$$T(1) = 1$$
$$T(2) = 2$$
$$T(3) = 3$$
$$T(4) = 4$$
$$T(5) = 5$$
$$T(6) = 6$$
$$T(7) = 7$$
$$T(8) = 8$$
$$T(9) = 9$$
$$T(10) = 10$$
$$T(11) = 11$$
$$T(12) = 12$$
$$f(1) = 9$$
$$f(2) = 46$$
$$f(3) = 48$$
$$f(4) = 23$$
$$f(5) = 15$$
$$f(6) = 24$$
$$f(7) = 8$$
$$f(8) = 15$$
$$f(9) = 7$$
$$f(10) = 4$$
$$f(11) = 5$$
$$f(12) = 4$$
$$f(13) = 29$$

Fig. 1 -- Input to the JOSS programs (data from Table 1)

Interval number	Lower limit	Upper limit	Observed freq.	Expected freq.
1	.00	1.00	9	40.73
2	1.00	2.00	46	33.73
3	2.00	3.00	48	27.93
4	3.00	4.00	23	23.13
5	4.00	5.00	15	19.16
6	5.00	6.00	24	15.87
7	6.00	7.00	8	13.14
8	7.00	8.00	15	10.88
9	8.00	9.00	7	9.01
10	9.00	10.00	4	7.46
11	10.00	11.00	5	6.18
12	11.00	12.00	4	5.12
13	12.00	inf.	29	24.67
		totals:	237	237.01

The value of chi-square is   55.53. There are   11 degrees of freedom.

The estimated occurrence rate is   .188544.
The estimated mean (= reciprocal of the rate) is   5.30.

The preceding test of exponentiality used the chi-square minimum method.

Fig. 2 -- JOSS output (turnaround time data from Table 1)

Example 2: For the data of Example 1, it is interesting to illustrate the dependence (cited at the end of Sec. 2.3) of chi-square tests upon the choice of the number and position of the intervals into which the time axis is divided. The data in Table 2 are the data of Table 1 arranged in three-hour intervals instead of one-hour intervals.

Table 2

NET TURNAROUND TIMES

Interval (Hours)	Frequency
0-3	103
3-6	62
6-9	30
9-12	13
more than 12	29
Total	237

The JOSS output for the chi-square minimum test of exponentiality for the data of Table 2 is given in Fig. 3. The data now appear to

Interval number	Lower limit	Upper limit	Observed freq.	Expected freq.
1	.00	3.00	103	101.12
2	3.00	6.00	62	57.97
3	6.00	9.00	30	33.24
4	9.00	12.00	13	19.06
5	12.00	inf.	29	25.61
		totals:	237	237.00

The value of chi-square is   3.00. There are   3 degrees of freedom.

The estimated occurrence rate is   .185418.
The estimated mean (= reciprocal of the rate) is   5.39.

The preceding test of exponentiality used the chi-square minimum method.

Fig. 3 -- JOSS output (turnaround time data from Table 2)

be in good agreement with an exponential distribution! Since the same data, differently grouped, were seen in Example 1 to be decidedly nonexponential, this shows how cautious one must be in *accepting* a hypothesis of exponentiality based on a chi-square test merely because that test *fails to reject* the hypothesis. Cf. Sec. 2.3.

<u>Example 3</u>: The data in Table 3 are times-to-failure of 118 AN/ARC-90 Radios, given by Allen and Sloan (1966).

The JOSS output for the chi-square minimum test of exponentiality is given in Fig. 4. The data seem to be in good agreement with an exponential distribution. As a further check, all observations above

Table 3

TIMES-TO-FAILURE

Interval (Hours)	Frequency
0-20	19
20-40	19
40-60	21
60-80	10
80-100	13
100-120	6
120-140	7
140-160	5
160-180	4
180-200	2
200-220	3
220-240	1
240-260	2
260-280	1
280-300	1
300-320	1
320-340	2
> 340	1
Total	118

Interval number	Lower limit	Upper limit	Observed freq.	Expected freq.
1	.00	20.00	19	23.10
2	20.00	40.00	19	18.58
3	40.00	60.00	21	14.94
4	60.00	80.00	10	12.02
5	80.00	100.00	13	9.66
6	100.00	120.00	6	7.77
7	120.00	140.00	7	6.25
8	140.00	160.00	5	5.03
9	160.00	180.00	4	4.04
10	180.00	200.00	2	3.25
11	200.00	220.00	3	2.62
12	220.00	240.00	1	2.10
13	240.00	260.00	2	1.69
14	260.00	280.00	1	1.36
15	280.00	300.00	1	1.09
16	300.00	320.00	1	.88
17	320.00	340.00	2	.71
18	340.00	inf.	1	2.91
		totals:	118	118.00

The value of chi-square is 10.08. There are 16 degrees of freedom.

The estimated occurrence rate is .010892.
The estimated mean (= reciprocal of the rate) is 91.81.

The preceding test of exponentiality used the chi-square minimum method.

Fig. 4 -- JOSS output (times-to-failure data from Table 3)

200 hours were combined into one class. The JOSS output resulting

from this accumulation is shown in Fig. 5. Again, good agreement

with the hypothesis of exponentiality is evident. The estimates of

the rates (and means) from the two different outputs differ somewhat

(as is to be expected), but not dramatically. There remains the

possibility of additional combinations of adjacent intervals to try

to achieve a significant departure from the exponential hypothesis

as discussed in Sec. 2.3. Further, the moral of Example 2 must not

be overlooked.

Example 4: To illustrate the approximation used with the modified

chi-square minimum method for testing exponentiality (discussed in

Sec. 3.1), consider the data of Table 3, rearranged as follows. (It

is known that the longest time-to-failure is 358 hrs.)

Table 4

TIMES-TO-FAILURE

Interval (Hours)	Frequency
0-20	19
20-40	19
40-60	21
60-80	10
80-100	13
100-120	6
120-140	7
140-160	5
160-180	4
180-200	2
200-360	12
> 360	0
Total	118

Applying Eq. (3.5) to the data of Table 4 yields β = 118/10,420 = .011324

and a calculated chi-square (from Eq. (2.1)) of 7.94. There are k = 12

Interval number	Lower limit	Upper limit	Observed freq.	Expected freq.
1	.00	20.00	19	23.91
2	20.00	40.00	19	19.07
3	40.00	60.00	21	15.20
4	60.00	80.00	10	12.12
5	80.00	100.00	13	9.67
6	100.00	120.00	6	7.71
7	120.00	140.00	7	6.15
8	140.00	160.00	5	4.90
9	160.00	180.00	4	3.91
10	180.00	200.00	2	3.12
11	200.00	inf.	12	12.26
		totals:	118	118.02

The value of chi-square is   5.65.   There are   9 degrees of freedom.

The estimated occurrence rate is   .011322.
The estimated mean (= reciprocal of the rate) is   88.32.

The preceding test of exponentiality used the chi-square minimum method.

Fig. 5 -- JOSS output (data from Table 3 with some combinations)

intervals (the last interval must be counted, even though it has a zero frequency), so this chi-square has 10 degrees of freedom. The calculated chi-square is decidedly nonsignificant.

The data of Table 4 were also analyzed by the chi-square minimum method and yielded an estimated rate of 0.011310 and a χ^2 of 7.94--the same value calculated by the approximate method of Sec. 3.1. This example illustrates that this approximation can be quite good. The discussion of Example 2 concerning the conclusion to draw from the analysis is pertinent to this example as well.

6. TESTING GROUPED DATA FOR FIT TO OTHER DISTRIBUTIONS THAN THE EXPONENTIAL

This study was undertaken in response to a request, and because the exponential is an important, widely used distribution. It turns out, however, that of all the commonly considered event-time distributions, it is simplest to test grouped data for fit to the exponential distribution. This stems from two reasons. First, the probability content of the intervals and the derivative of this probability with respect to the parameter can be given in a closed form expression. No table look-up or numerical approximation is needed. Second, only one parameter is involved, so that the chi-square expression needs to be minimized as a function of only one variable—the rate parameter. One-dimensional minimization procedures are, of course, easier to implement than multi-dimensional procedures.

Some distributions commonly used to describe event-times are listed below along with a brief discussion of the problems involved in testing grouped data for fit to each one. In each case, I denotes the interval $L < t < U$.

(1) <u>The Weibull Distribution</u>. The Weibull distribution has density

$$f(t) = \lambda \alpha t^{\alpha-1} \exp(-\lambda t^{\alpha}) , \quad t \geq 0 . \tag{6.1}$$

The probability content of the interval I, as a function of the parameters λ and α, is

$$p_I(\lambda,\alpha) = \exp(-\lambda L^\alpha) - \exp(-\lambda U^\alpha) \quad . \tag{6.2}$$

The partial derivatives of this probability with respect to λ and α are

$$\frac{\partial p_I(\lambda,\alpha)}{\partial \alpha} = U^\alpha \exp(-\lambda U^\alpha) - L^\alpha \exp(-\lambda L^\alpha) \tag{6.3}$$

and

$$\frac{\partial p_I(\lambda,\alpha)}{\partial \alpha} = \lambda U^\alpha (\log U) \exp(-\lambda U^\alpha) - \lambda L^\alpha (\log L) \exp(-\lambda L^\alpha) \quad . \tag{6.4}$$

For the Weibull, Eqs. (2.5) become two (nonlinear) equations in the two unknowns λ and α. One method of solving this pair of equations is with a two-dimensional analog of the Newton-Raphson method. See Ralston (1965, Sec. 8.8).

Thus, for the Weibull, the probability content of the intervals and the derivatives of this probability with respect to the parameters all can be given as closed form expressions. However, there is a two-dimensional minimization problem to solve.

(ii) The Gamma Distribution. The gamma distribution has density

$$f(t) = \lambda^\alpha t^{\alpha-1} e^{-\lambda t}/\Gamma(\alpha) \quad , \quad t \geq 0 . \tag{6.5}$$

The probability content of the interval I, as a function of the parameters λ and α, is

$$p_I(\lambda,\alpha) = \int_{\lambda L}^{\lambda U} [z^{\alpha-1} e^{-z}/\Gamma(\alpha)] \, dz \quad , \tag{6.6}$$

a quantity related to the incomplete gamma function. This function is

well tabulated and good algorithms exist for computing it. The partial derivatives of this probability with respect to λ and α are

$$\frac{\partial p_I(\lambda,\alpha)}{\partial \lambda} = \lambda^{\alpha-1} \left[U^\alpha e^{-\lambda U} - L^\alpha e^{-\lambda L} \right] / \Gamma(\alpha) \qquad (6.7)$$

and

$$\frac{\partial p_I(\lambda,\alpha)}{\partial \alpha} = \int_{\lambda L}^{\lambda U} \{ z^{\alpha-1} e^{-z} [\Gamma'(\alpha) - \Gamma(a)\log z]/\Gamma^2(\alpha) \} \, dz \quad . \qquad (6.8)$$

The partial derivative in Eq. (6.8) is a fairly complicated quantity to calculate and this suggests that rather than trying to solve the pair of equations corresponding to Eqs. (2.5), it may be preferable to attempt a direct numerical minimization of the chi-square expression (2.4). See, for example, Wilde (1964).

Thus, for the gamma distrubution, one is faced with complicated expressions to evaluate--or rather, approximate, and a two-dimensional minimization problem.

(iii) The Truncated Normal Distribution. The truncated normal distribution is the ordinary normal distribution truncated on the left at zero so that it is the distribution function of a nonnegative random variable. Its density is given by

$$f(t) = (a\sigma)^{-1}(2\pi)^{-\frac{1}{2}} \exp[-(t-\mu)^2/(2\sigma^2)] \;,\; t \geq 0 \qquad (6.9)$$

where a, a function of μ and σ, is given by

$$a = 1 - \Phi(-\mu/\sigma) \quad , \qquad (6.10)$$

with $\Phi(x)$ being the cumulative distribution function for the standard normal distribution

$$\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-\frac{1}{2}} \exp(-t^2/2) \, dt \; . \tag{6.11}$$

The probability content of the interval I, as a function of the parameters μ and σ is

$$p_I(\mu,\sigma) = [\Phi((U-\mu)/\sigma) - \Phi((L-\mu)/\sigma)]/a \; . \tag{6.12}$$

Interested readers are invited to calculate the partial derivatives of $p_I(\mu,\sigma)$ with respect to μ and σ and to determine whether the pair of equations corresponding to (2.5) should be solved, or whether a direct numerical minimization of Eq. (2.4) is preferable here. The same might be done for the lognormal distribution.

(iv) The Lognormal Distribution. The lognormal distribution has density function

$$f(t) = (t\sigma)^{-1}(2\pi)^{-\frac{1}{2}} \exp[-(\log t - \mu)^2/(2\sigma^2)] \; , \; t > 0 \; . \tag{6.13}$$

The probability content of the interval I, as a function of the parameters μ and σ, is

$$p_I(\mu,\sigma) = \Phi((\log U-\mu)/\sigma) - \Phi((\log L-\mu)/\sigma) \; . \tag{6.14}$$

7. TESTING NONGROUPED DATA FOR EXPONENTIALITY

If one has individual event times t_1, ..., t_n, the chi-square minimum method discussed here is not the best for testing whether these data conform to an exponential distribution. Reasons for this have already been discussed in Sec. 2.3. In particular, the chi-square procedure tests the hypothesis of exponentiality against an unrestricted alternative. One may want to test for exponentiality against some specified alternative. Epstein (1960) discusses a number of procedures for this. Some subsequent papers are Proschan and Pyke (1964), Jackson (1967), and Barlow (1968).

A very simple, though qualitative, test of exponentiality is the following graphical procedure. Since the survival probability function for the exponential distribution with rate ρ is $\exp(-\rho t)$, plotting the observed failures t_i against the negative of the logarithm of the empirical survival probability function should yield pretty close to a straight line if the hypothesis of exponentiality holds. Actually, Epstein recommends plotting t_i against $-\log\left(\frac{n-i+1}{n+1}\right) = \log\left(\frac{n+1}{n-i+1}\right)$ because the expected value of $\bar{F}(t_i)$ is $\frac{n-i+1}{n+1}$. [Here $\bar{F}$ denotes the survival probability function and t_i the i-th *ordered* observation.] This procedure is valid even for censored samples (in which only the first r event times out of the n times being observed are available) and for truncated samples (in which observation ceases at a predetermined time T).

Epstein, incidentally, makes a misleading statement in connection with the chi-square test of exponentiality. He states that one should use "the best estimate" of the parameter in our expression (2.1) and

that the result is then distributed as chi-square with k-1 degrees of freedom. Although Epstein does not state what he means by "best," it is known that if one uses in (2.1) the maximum likelihood estimate of the parameter based on the individual observations, then the limiting distribution, under suitable regularity conditions, lies between (in the sense of stochastic ordering) the chi-square distributions with k-1 and k-2 degrees of freedom.

This follows from a result of Chernoff and Lehmann (1954) to the following effect:

If individual observations t_1, ..., t_n are available and if one substitutes into Eq. (2.4) the maximum likelihood estimates $\hat{\alpha}_1$, ..., $\hat{\alpha}_s$ of α_1, ..., α_s, then under suitable regularity conditions the asymptotic distribution of $\chi^2(\hat{\alpha}_1, ..., \hat{\alpha}_s)$ lies (in the sense of stochastic ordering) between the chi-square distributions with k-1 and k-s-1 degrees of freedom.

For the exponential distribution there is one parameter, i.e., s = 1.

Thus, if one follows Epstein's advice (interpreting his "best" to mean maximum likelihood), he will reject the null hypothesis of exponentiality (when it is actually true) less often than he should. That is, one may think he has a test with probability of type I error equal to, say, 0.05, but it is really smaller.

Appendix

JOSS PROGRAM LISTING

1.1 Demand k.
1.2 Do part 2 for j=1(1)(k-1).
1.3 Set T(0)=0.
1.4 Do part 3 for j=1(1)k.
1.5 To part 4.

2.1 Demand T(j).

3.1 Demand f(j).

4.1 Set D=sum[j=1(1)(k-1): f(j)·(X(j-1)+X(j))/2] + f(k)·X(k-1).
4.2 Do part 5 for r=n/D.

5.0 Set R=t(r).
5.1 Page.
5.2 Type form 1 if $=1.
5.3 Type form 2 if $=2.
5.4 Do part 6 for i=1(1)(k-1).
5.5 Type k,T(k-1),f(k),v(k,R) in form 4.
5.6 Line.
5.7 Type n,sum[i=1(1)k:v(i,R)] in form 5.
5.71 Line.
5.8 Type G(R),(k-2) in form 8.
5.81 Line.
5.9 Type R/T(k-1) in form 6.
5.91 Type T(k-1)/R in form 7.
5.92 Line.
5.93 Type form 9.

6.1 Line if $=3.
6.2 Type i,T(i-1),T(i),f(i),v(i,R) in form 3.

Form 1:
Interval   Lower   Upper   Observed   Expected

Form 2:
 number   limit   limit   freq.   freq.

Form 3:

 —   ——·—— ——·——   ——   ——·——

Form 4:

 —   ——·—— inf.   ——   ——·——

Form 5:

              totals: ——   ——·——

Form 6:
The estimated occurrence rate is __._____.

Form 7:
The estimated mean (= reciprocal of the rate) is _____.__.

Form 8:
The value of chi-square is ___.__. There are ___ degrees of freedom.

Form 9:
The preceding test of exponentiality used the chi-square minimum method.

```
    G(r):   sum[i=1(1)k: ((f(i)-n•p(i,r))*2)/(n•p(i,r))]
    I(r):   r-M(r)/m(r)
    M(r):   sum[i=1(1)k: [h(i,r)+h(i,r)*2/(2•n)]•q(i,r)]
    X(i):   T(i)/T(k-1)
      d:    10*(-4)
  e(i,r):   exp(-r•X(i))
  h(i,r):   f(i)/p(i,r) - n
    m(r):   [M(r+d)-M(r)]/d
      n:    sum[j=1(1)k: f(j)]
  p(i,r):   [i=k: e(k-1,r); e(i-1,r)-e(i,r)]
  q(i,r):   [i=k: -X(k-1)∘e(k-1,r); X(i)•e(i,r)-X(i-1)•e(i-1,r)]
    t(r):   [|M(r)|<10*(-6): r; t(I(r))]
  v(i,r):   ip(100•n•p(i,r)+.5)/100
```

REFERENCES

1. Allen, J., and H. J. Sloan (1966). *Reliability of Airborne Electronic Equipment and Its Effect on Logistics Requirements.* Air Force Institute of Technology, Master's Thesis No. SLSR-35-66.

2. Barlow, R. E. (1968). Likelihood ratio tests for restricted families of probability distributions. *Ann. Math. Statist., 39*, 547-560.

3. Berkson, J. (1966). Examination of randomness of α-particle emissions in F. N. David (ed.), *Festschrift for J. Neyman*, John Wiley and Sons, Inc., New York, 37-54.

4. Bryan, G. E., and E. W. Paxson (1967). *The JOSS Notebook.* The RAND Corporation, RM-5367-PR.

5. Chernoff, Herman, and E. L. Lehmann (1954). The use of the maximum likelihood estimate in χ^2 tests for goodness of fit. *Ann. Math. Statist., 25,* 579-586.

6. Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10,* 417-451.

7. Cramér, Harald (1951). *Mathematical Methods of Statistics,* Princeton University Press, Princeton.

8. Epstein, Benjamin (1960). Tests for the validity of the assumption that the underlying distribution of life is exponential. *Technometrics, 2,* 83-101 and 167-183.

9. Jackson, O.A.Y. (1967). An analysis of departures from the exponential distribution. *J. Roy. Statist. Soc. Ser. B, 29,* 540-549.

10. Proschan, F., and R. Pyke (1964). Asymptotic normality of certain test statistics of exponentiality. *Biometrika, 51,* 253-255.

11. Ralston, Anthony (1965). *A First Course in Numerical Analysis,* McGraw-Hill Book Company, New York.

12. Wilde, D. J. (1964). *Optimum Seeking Methods.* Prentice-Hall, Inc., Englewood Cliffs, N.J.

DOCUMENT CONTROL DATA

1. ORIGINATING ACTIVITY	2a. REPORT SECURITY CLASSIFICATION
THE RAND CORPORATION	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

TESTING GROUPED DATA FOR EXPONENTIALITY

4. AUTHOR(S) (Last name, first name, initial)

Scheuer, Ernest M.

5. REPORT DATE	6a. TOTAL No. OF PAGES	6b. No. OF REFS.
August 1968	40	12

7. CONTRACT OR GRANT No.	8. ORIGINATOR'S REPORT No.
F44620-67-C-0045	RM-5692-PR

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
DDC-1	United States Air Force Project RAND

10. ABSTRACT	11. KEY WORDS
A treatment of the problem of testing grouped data for fit to an exponential distribution. A JOSS computer program is provided to implement an appropriate test (the chi-square minimum method), and examples are given to show how to use this program. There is also some discussion of testing nongrouped data for exponentiality and of testing grouped data for fit to certain nonexponential distributions.	Statistical methods and processes Testing Data processing Computer programs JOSS Probability Reliability